

Figure 6: **Qualitative comparisons with MoGe [60].** We provide qualitative comparisons with the concurrent work MoGe [60]. Top: input images are taken from four test sets: Hypersim [44], DIODE [56], ScanNet [8], and ETH3D [50]. Middle: results of MoGe [60]. Bottom: our results. As a discriminative model, MoGe [60], like previous discriminative models [68, 4], also suffers from *flying pixels* at edges and details.

Table 5: **Quantitative comparisons with REPA [76].** Our model significantly outperforms REPA [76]. To ensure a fair comparison, the pretrained vision encoder used in both DiT+REPA and DiT+Ours is kept the same.

Method	NYUv2		KITTI		ETH3D		ScanNet		DIODE	
	AbsRel↓	$\delta_1$ ↑	AbsRel↓	$\delta_1$ ↑	AbsRel↓	$\delta_1$ ↑	AbsRel↓	$\delta_1$ ↑	AbsRel↓	$\delta_1$ ↑
DiT (baseline)	22.5	72.8	27.3	63.9	12.1	87.4	25.7	65.1	23.9	76.5
DiT+REPA [76]	17.6	78.0	23.4	70.6	9.1	91.2	20.1	74.3	14.6	86.9
DiT+Ours	<b>4.3</b>	<b>97.4</b>	<b>8.0</b>	<b>93.1</b>	<b>4.5</b>	<b>97.7</b>	<b>4.5</b>	<b>97.3</b>	<b>7.0</b>	<b>95.5</b>

## A More Qualitative Comparisons

We provide qualitative comparisons with the concurrent work MoGe [60], as shown in Figure 6. MoGe [60], as a discriminative model, suffers from *flying pixels* at edges and fine structures, a common issue observed in other discriminative models [68, 4]. Our model produces significantly fewer flying pixels compared to MoGe [60].

## B Additional Discussion with REPA

We provide an additional discussion on the recent image generation method REPA [76]. REPA [76] aligns intermediate tokens in diffusion models with pretrained vision encoder, significantly improving training efficiency and generation quality for image generation tasks. We compare our method with REPA [76], and the quantitative evaluation results are presented in Table 5. DiT+REPA refers to training the DiT model with REPA’s representation alignment, while DiT+Ours denotes training the DiT model using our Semantics-Guided DiT. For a fair comparison, the pretrained vision encoder used in both DiT+REPA and DiT+Ours is kept the same. Experimental results show that our Semantics-Guided DiT significantly outperforms REPA [76]. We attribute our model’s superiority over REPA to two factors. First, during training, REPA’s implicit alignment of DiT tokens with the pretrained vision encoder is suboptimal, making it difficult for DiT to effectively leverage semantic guidance from the pretrained vision encoder. In contrast, our Semantics-Guided DiT directly integrates semantic cues, resulting in more effective guidance. Second, at inference, REPA cannot leverage the pretrained vision encoder to provide semantic guidance, whereas our method effectively incorporates high-level semantics into the Semantics-Guided DiT during inference to guide the diffusion process.

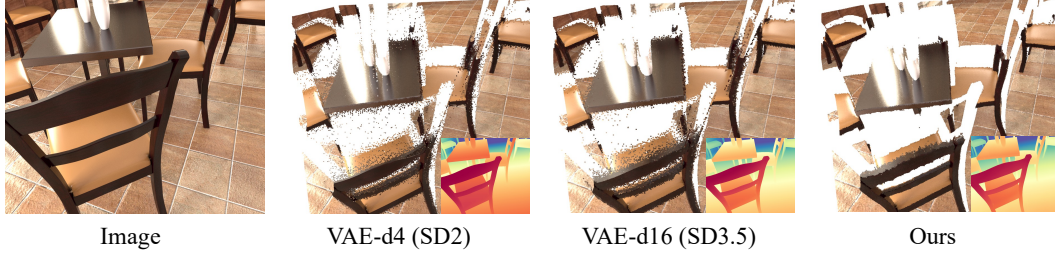


Figure 7: **Validation of flying pixels in different types of VAEs.** We present further qualitative comparisons showing that increasing the latent dimension in VAEs fails to eliminate *flying pixels*. VAE-d4 (SD2) denotes the reconstruction of ground truth depth maps using the VAE from Stable Diffusion 2, with a latent dimension of 4, which is also used in Marigold. VAE-d16 (SD3.5) uses the VAE from Stable Diffusion 3.5, which has a latent dimension of 16.

## 523 C Analysis of Flying Pixels in Different Types of VAEs

524 To better understand the emergence of *flying pixels* in VAE-based reconstructions, we analyze VAEs  
 525 with different latent dimensions (*i.e.*, channel) by using them to reconstruct ground truth depth maps.  
 526 Figure 7 shows that both VAE variants exhibit flying pixels at object edges and details, revealing a  
 527 common weakness of VAE reconstructions in preserving precise geometric structures. VAE-d4 (SD2)  
 528 denotes the reconstruction of ground truth depth maps using the VAE from Stable Diffusion 2, with a  
 529 latent dimension of 4, which is also used in Marigold [29]. VAE-d16 (SD3.5) uses the VAE from  
 530 Stable Diffusion 3.5, which has a latent dimension of 16.